Check for updates

Journal of KIIT. Vol. 21, No. 10, pp. 69-82, Oct. 31, 2023. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2023.21.10.69

에세이 창의성 점수 예측 모델을 위한 학습데이터 구축 방안

김덕기*[1], 한상우*[2], 박승혁*[3], 김용연*[4], 박상현*[5], 조준영*[6], 최서인*[7], 이유빈*[8], 유대곤**, 온병원***

Training Set Creation Method for Essay Creativity Score Prediction Models

Deokgi Kim*[1], Sangwoo Han*[2], Seunghyeok Park*[3], Yougyeon Kim*[4], Sanghyun Park*[5], Junyoung Jo*[6], Seoin Choi*[7], Youbin Lee*[8], Daegon Yu**, and Byung-Won On***

이 논문은 2022년도 정부(과기정통부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(2022R1A2C1011404)

요 약

학생들은 에세이를 작성해보면서, 제4차 산업혁명에 꼭 필요한 창의성을 키울 수 있다. 그러나 현재 수작업에 의존하는 창의성 평가방식은 주관적이며, 많은 예산과 시간이 소요된다. 따라서 대량의 에세이에서 창의성을 정량적으로 평가할 수 있도록 인공지능 모델을 개발하는 것이 중요하다. 본 논문에서 에세이 창의성 점수 예측 모델에 필요한 학습데이터를 구축하는 방안을 제안하고, 대용량 학습데이터를 효과적으로 구축할 수 있도록 웹 기반 레이블링 시스템을 개발한다. 제안 시스템을 통해, 약 3,766개의 에세이에 대한 학습데이터를 구축하였고 심도 있는 분석을 통해, 3명의 평가자 간 Kendall 점수는 평균 82.7%로 평가자 간 강한 상관관계를 보였고, 가설검정을 통해 제안 시스템을 사용하는 것이 기존 수작업 방식보다 통계적으로 유의미한 것을 보였다.

Abstract

By writing essays, students can develop creativity, which is essential for the 4th Industrial Revolution. However, the current creativity evaluation method that relies on manual work is subjective and requires a lot of budget and time. Therefore, it is important to develop an artificial intelligence model to quantitatively evaluate creativity in large-scale essays. In this paper, we propose a method to build the training set needed for the essay creativity score prediction model and develop a web-based labeling system to effectively build large amounts of training set. Through the proposed system, training set for approximately 3,766 essays was constructed and through in-depth analysis, the Kendall score between the three evaluators showed a strong correlation between evaluators with an average of 82.7%, and through hypothesis testing the proposed system was statistically more significant than the existing manual method.

Keywords

automated essay creativity scoring, creativity assessment, computational creativity, training set creation method, ASAP

* 군산대학교 소프트웨어학부
- ORCID[1]: https://orcid.org/0000-0002-8429-6261
- ORCID[2]: https://orcid.org/0000-0002-7514-2041
- ORCID[3]: https://orcid.org/0009-0000-3844-4270
- ORCID[4]: https://orcid.org/0000-0002-1502-1205
- ORCID[5]: https://orcid.org/0009-0005-9826-7815
- ORCID[6]: https://orcid.org/0000-0002-9944-2401
- ORCID[7]: https://orcid.org/0009-0001-3913-0940
- ORCID[8]: https://orcid.org/0000-0002-2596-6990

** ㈜애니파이브 연구원
- ORCID: https://orcid.org/0000-0003-4331-8302

*** 군산대학교 소프트웨어학부 교수(교신저자)
- ORCID: https://orcid.org/0000-0001-6929-3188

· Received: Sep. 10, 2023, Revised: Oct. 16, 2023, Accepted: Oct. 19, 2023

· Corresponding Author: Byung-Won On
School of Software, Kunsan National University, 558, Daehak-ro, Gunsan, Jeollabuk-do, Korea
Tel.: +82-63-469-8913, Email: bwon@kunsan.ac.kr

Ⅰ. 서 론

제4차 산업혁명 시대에는 기존의 지식을 바탕으로 새로운 지식을 창출하는 능력이 강조되고 있다. 특히 제4차 산업혁명의 핵심인 인공지능과의 협업을 통해 인간은 반복적인 작업을 넘어서 창의적인 활동에 집중하여 혁신을 이루어낼 수 있다. 이러한 지식 창출은 인간의 창의적인 사고를 기반으로 이루어지며[1], 제4차 산업혁명의 핵심적인 요소로 작용한다.

이러한 창의성은 학문 분야에 따라 다양한 면을 갖고 있는데, 언어학, 심리학, 문학 등 각 학문에서의 창의성의 본질은 조금씩 달라진다. 언어학적 관점에서는 한정된 언어 자원을 활용하여 문장을 창출하는 능력을 강조하며, 심리학적 관점에서는 언어 활동과 인지심리학적 요소를 중심으로 창의성을 이해한다. 문학적 관점에서는 관습에서 벗어나는 정도를 창의성의 평가 기준으로 활용하며, 이러한 관점들이 창의성의 다면적인 특성을 드러내고 있다. 이처럼 창의성은 그 정의가 복잡하고 다양하여 학술적으로도 명확한 정의를 내리기가 어렵다.

그런데도 여러 관점에서 다양한 정의를 제시했다. [2]는 “문제, 결함, 지식의 공백, 부조화에 민감한 과정”으로, [3]은 “새롭고 유용한 아이디어의 생성”으로, [4]는 “개인이나 집단이 사회적 맥락에서 새롭고 유용한 결과를 창출하는 능력과 과정”으로 창의성을 정의했다. 현재까지 가장 널리 수용되는 정의는 “새롭고 적절한 것을 창조하는 능력”으로 의견이 모이고 있다.

창의적인 텍스트는 새로우면서 동시에 독자, 즉 사회·문화적 맥락 내에서 소통할 수 있는 텍스트를 말하며, 창의적인 글쓰기(Creative writing)는 창의적인 텍스트를 작성하는 행위로 필자가 자신의 새롭고 독창적인 아이디어를 사회·문화적 맥락 내에 적절하고 효과적인 텍스트로 소통될 수 있도록 표현하는 글쓰기 활동을 의미한다. 이처럼, 글쓰기를 통해 중고등학교 학생들은 제4차 산업혁명에 꼭 필요한 창의성을 키울 수 있는 능력을 기르게 된다.

특히 에세이는 글쓰기의 한 형태로, 학생들은 주제를 깊이 있게 이해하고 그것을 구조화된 글로 표현하는 데 노력하며 자기 생각과 견해를 전달한다.

그러나 현재의 에세이 창의성 평가방식은 주관적인 평가에 크게 의존하고 있다. 이에 따라 신뢰성 있는 결과를 얻기 어려우며, 평가자 간의 의견이 일치하는 경우가 드물다. 이러한 문제를 해결하기 위해 평가자들은 협의 과정을 통해 개인의 평가 결과를 보완하여 신뢰도를 높일 수 있지만, 많은 시간이 소요되고 전문가 초빙으로 큰 비용이 소요된다. 더구나, 수백 또는 수천 건의 에세이를 상세히 검토하고 평가하는 것은 현실적으로 어려운 일이다. 이와 관련하여 [5]의 연구 결과에 따르면, 다양한 평가자들의 의견은 평가 결과에 큰 차이를 불러오며, 창의성 평가의 객관성을 확보하기 위해서는 더 깊은 연구가 필요함을 시사하고 있으며, 현재 에세이의 창의성을 평가하기 위한 적절한 학습데이터가 구축되지 않았다. 이에 따라, 본 논문은 주관적이지 않고, 객관적이며 정량화된 방법을 통해 에세이의 창의성을 평가할 수 있는 학습데이터를 체계적으로 구축하는 방안에 관해 기술한다.

본 연구의 기여도는 다음과 같다.

- 인공지능 모델 학습을 위한 창의적인 에세이 학습데이터를 구축하는 방안을 처음으로 제안하고, 약 3,766개의 에세이가 포함된 학습데이터를 구축했다. 생성된 학습데이터를 분석한 결과, 3명의 평가자 간 평균 Kendall 점수는 약 82.7%로 평가자 간의 강한 상관관계를 보였다.
- 대용량 학습데이터를 효율적으로 생성하고 검증하기 위해 웹 기반 레이블링 시스템을 제안하고 학습데이터 구축에 활용하였다. 가설검정을 통해 제안하는 시스템은 전통적인 평가 방법과 비교하여 통계적으로 유의미한 것으로 나타났다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구에 관해 정리하고 3장에서는 학습데이터 구축 방안에 관해 기술한다. 4장에서는 구축한 학습데이터의 통계에 관해 설명하고, 5장에서는 실험 결과에 대해 자세히 논의한다. 마지막으로 6장에서는 본 논문의 결론 및 향후 연구를 기술한다.

Ⅱ. 관련 연구

공학 분야에서 다양한 창의성 관련 연구와 해결

방안이 제시됐다. 초기에는 텍스트 데이터의 참신성 탐지를 위한 상대 문서 벡터 기반의 합성신경망 모델이 제안되었다[6]. 그 후로, 논문의 참신성을 탐지하기 위한 기계학습 모델[7], 텍스트 데이터의 참신성을 접속사절 기반의 체틀린머신을 이용하여 탐지하려는 모델[8], 참신한 아이디어를 탐지하기 위한 휴리스틱 모델이 있다[9]. 최근에는, 휴렛패커드 데이터 세트를 활용하여 창의적인 에세이를 추천하기 위한 적대적 생성 신경망 모델 등이 제안됐다[10].

이처럼 대부분의 연구는 학생 에세이의 창의성 평가보다는 논문, 특허, 디자인, 이미지 등 특정 도메인에 필요한 “참신성 탐지(Novelty Detection)”에 국한되어 있고 소량의 학습데이터를 활용하였다. 이에 반해, 본 연구는 글쓰기 수업 현장에서 학생들이 작성한 에세이를 신속히 평가하고 빠른 피드백을 제공함으로써 교사와 학생 간 수업의 질을 향상하는 데 도움을 줄 수 있는 인공지능 모델 개발과 관련되어 있다. 특히, 인공지능 모델 학습을 위해 필요한 에세이 창의성 평가 학습데이터를 구축하기 위해, 국어·교육학·심리학 분야에서 오랫동안 연구해왔던 이론적인 창의성 평가 지표를 적절히 적용하여 창의성 평가를 위한 학습데이터를 구축하고, 품질을 검증한다. 또한, 향후 대용량 창의성 학습데이터를 구축할 수 있도록 웹 기반의 레이블링 시스템을 개발하고 정량적·정성적인 평가를 수행하여 제안 시스템의 우수성을 입증한다.

III. 학습데이터 구축 방안

그림 1은 에세이 창의성 평가를 위한 학습데이터 구축 과정을 도식화한 것이다.

3.1 에세이 글 데이터 번역 및 수정

본 논문에서는 에세이 점수를 예측하는 인공지능 모델을 학습하는 데 많이 활용되고 있는 ASAP(Automated Student Assessment Prize)라는 벤치마크 데이터 세트[11]를 원시데이터로 사용하였다. 주제는 ‘도서관 검열’이며, 종류는 논설문으로 총 1,800편이다. ASAP는 영어로 작성되었기 때문에 한국어로 번역하기 위해 구글 번역기를 사용하였다[12]. 이 과정에서 문법이나 표현이 어색한 문장들은 [13]를 참고하여 다음과 같이 수정하였다.

1) 주어와 술어의 호응: 문장 내에서 주어와 해당 주어에 맞는 술어가 일치하는 문법적인 원칙을 의미한다. 이는 문장의 구조와 논리적 흐름을 유지하기 위해 중요한 요소이다. 예를 들어, “저의 장점은 남을 잘 돕는다고 생각합니다”라는 문장을 “저의 장점은 남을 잘 돕는 것입니다”로 더 자연스럽게 수정할 수 있다.

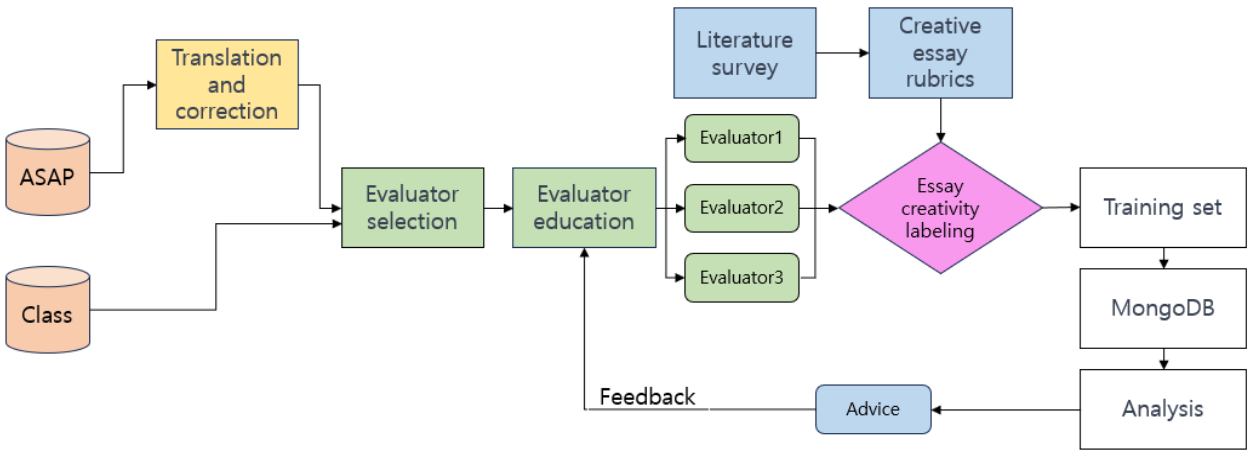


그림 1. 에세이 창의성 평가를 위한 학습데이터 구축 과정
Fig. 1. The overall process of constructing a training set for the essay creativity score prediction model

2) 잘못된 조사 사용 여부: 조사는 언어에서 단어와 단어 사이, 또는 문장 내에서 단어와 문장 간의 관계를 명확하게 표현하고 조정하는 역할을 하는 문법적 요소이다. 예를 들어, “저는 친구와 영화에 가고 싶어요”라는 문장은 “저는 친구와 영화를 보러 가고 싶어요”로 수정되어야 한다.

3) 대명 형용사: 명사와 형용사의 기능을 모두 가지면서, 대명사처럼 다른 명사를 대신하여 사용할 수 있는 품사를 의미한다. 예를 들어, “한국 청년의 역량은 그들이 나이를 먹을수록 감소한다. OECD 주관으로 국가별 문제 해결력을 조사한 PIAAC 데이터에 의하면 그러하다”라는 문장은 ‘그들이’라는 불필요한 대명 형용사를 사용하여 두 문장으로 나누었다. 이를 제거하여, “OECD가 주관하여 국가별 문제 해결력을 조사한 PIACC 자료를 보면, 한국 청년은 나이가 증가할수록 그 역량이 감소한다”와 같이 더 자연스럽게 수정할 수 있다.

4) 중복표현: 같거나 유사한 뜻을 가진 단어나 구를 중복해서 사용하는 문체적인 스타일을 말한다. 예를 들어, “강풍에 가로수 나무가 쓰러졌다”에서 ‘가로수’와 ‘나무’는 중복표현이다. 따라서 “강풍에 가로수가 쓰러졌다”라는 문장으로 수정해야 한다.

5) 긴 문장: 중복표현을 여러 개 사용하여 길어진 문장을 말한다. 예를 들어, “쓸데없이 세월을 낭비하고 헛되이 시간을 보내던”을 “세월을 낭비하던”으로 수정할 수 있다.

6) 잘 읽히지 않는 문장: “흔히 인간의 행동에는 정신이 선행되어야 하며 그렇기에 행동은 정신 과정의 결과로 생각하곤 한다”라는 문장은 “흔히 인간의 정신이 행동의 원인이라고 생각한다”처럼 읽기 쉽게 수정할 수 있다.

7) 직역투 문장: 영어, 일본어 등 다른 언어를 그대로 직역한 듯한 문장을 말한다. 예를 들어, “In general, studying is difficult to find interesting”을 직역하면 “일반적으로 공부는 재미를 찾기가 어렵습니다”로 된다. 이러한 직역투는 “일반적으로 공부는 재미없다”처럼 우리말 표현으로 자연스럽게 수정할 수 있다.

8) 대명사: 실제 명사를 대신하여 사용되는 품사를 의미한다. 예를 들어, “만수는 어제 학교에서 친구 철수와 싸웠다. 철수가 만수의 스마트폰을 몰래 훔쳐봤기 때문이다”라는 문장에서 만수는 ‘그는’으로, 철수는 ‘그의 친구’ 혹은 ‘친구’라는 지시대명사로 고정하여 “만수는 어제 학교에서 친구 철수와 싸웠다. 그의 친구가 그의 스마트폰을 몰래 훔쳐봤기 때문이다”로 수정할 수 있다.

9) 접속사: 문장 구조와 의미 전달에 중요한 역할을 하므로, 적절하게 사용하는 것이 중요하다. 접속사 없이 앞뒤 문장을 연결할 때 많은 글이 모호한 상황을 대명사로 표현하는 경우가 많다. 예를 들어, “평가 과정에서 그 내용을 의식적으로 인지하기 이전에 평가자에게 기저 요인들이 평정에 영향을 미칠 수 있다. 이는 평가자가 피평가자에게 가지는 감정적인 요인들을 주로 포함한다”라는 문장에서 ‘이는’이라는 대명사 표현이 무엇을 가리키는지가 명확하지 않다. 대명사 대신 ‘그런’이라는 접속사를 사용하여 “평가자가 의식적으로 인지하지 못하는 요인들이 평가에 영향을 줄 수 있다. 그런 요인 중 하나는 평가자의 피평가자에 대한 감정이다”로 수정해야 한다.

10) 시제 일치: 중문 혹은 복문은 시제가 일치하는지도 검토해야 한다. “나는 그가 일찍 일어난다고 들었다”라는 문장을 과거형으로 시제를 일치하여 “나는 그가 일찍 일어났다고 들었다”라고 수정할 수 있다.

11) 일본식 표현: 명사 뒤에 붙은 ‘~의’나 ‘~적’, 복수를 뜻하는 ‘들’, ‘하는 것’ 같은 일본식 표현에 주의해야 한다. 예를 들어, ‘가족과의 솔직한 소통’은 ‘가족과 솔직하게 소통하기’로, ‘사회적 문제’는 ‘사회 문제’로, ‘많은 관광객이 거리의 위험들에 노출되어’라는 문장은 ‘많은 관광객이 거리의 여러 위험에 노출되어’로, “글을 쓰는 것은 곧 생각을 정리하는 것이다”라는 문장은 “글쓰기는 생각을 정리할 수 있게 한다”로 수정할 수 있다.

이러한 검토 및 수정 작업을 마친 한국어 ASAP 데이터 세트와 기초글쓰기 수업에서 한국 학생들이 작성한 에세이를 원천데이터로 사용하였다. 표 1은 원천데이터 통계이다. 영어 및 한국어 ASAP 3,600편과 기초글쓰기 에세이 166편으로 총 3,766편이며, 종류별로 논설문은 3,664편과 정보글은 102편이다.

표 1. 에세이 데이터 통계
Table 1. Essay data characteristics

Prompt	Essay topic	# of essays	Type
1	Library censorship	3,600	Editorial article
2	Will artificial intelligence technology benefit or harm future humans?	64	
3	My success or failure story	38	Information article
4	Introducing my favorite cultural content (YouTube, Webcomics, Music, etc.)	16	
5	A day without smart devices	6	
6	What I'm doing and can do in the age of environmental pollution	9	
7	Python programming language	33	
Total		3,766	

3.2 에세이 글 데이터 평가

표 2는 논설문에 대한 창의성 평가 기준[14]을 정리한 표이다. 현재까지 글쓰기 평가에서의 창의성은 문학적인 글을 작성하는 데 초점을 맞추어 왔다.

소설 또는 시와 같은 문학 작품을 창작할 수 있는가를 가지고 창의성을 측정하는 기준으로 삼았지만, 실생활에서 가장 많이 사용되고 있는 정보글이나 논설문 같은 실용적인 글쓰기에 대한 논의는 제대로 이루어지지 않고 있다[15]. 자기 생각을 사회적이고 문화적인 맥락에 적합하도록 창의적으로 표현하는 글쓰기 훈련이 필요하며, 글쓰기 평가도 논설문이나 정보글의 창의성을 제대로 평가할 수 있어야 한다.

국어 교육학 및 심리학 분야에서는 주로 학생이 작성한 문학 작품에서 창의성을 찾고 이를 점수화하기 위해 다양한 평가 지표를 개발하였다[5][14][16]. 특히, [14]의 연구는 기존의 창의성 평가가 개인의 주관적인 평가와 평가자의 협의와 같은 전통 방식에서 벗어나, 보다 객관적으로 측정할 수 있는 평가 방법을 제시하였다.

본 논문에서는 원천데이터에 대해 표 2의 9가지 평가 요소에 해당하는 평가 기준으로 1~5점으로 평가하였다.

표 2. 에세이 창의성 평가 지표
Table 2. Essay creativity rubrics

Category	Evaluation index	Evaluation criteria	Score range
Content(C)	Novelty of ideas and content(C1)	The idea is ingenious. The argument or evidence is fresh and novel.	1~5
	Richness of content(C2)	The content (evidence, examples, etc.) is diverse. The content is specific.	1~5
	Logicality of content(C3)	The evidence is valid and reasonable. The author responds appropriately, taking into account expected counterarguments.	1~5
Organization(O)	Originality of structure(O1)	The narrative approach is unique. The introduction and conclusion have been impressively structured.	1~5
	Cohesiveness of structure(O2)	Paragraphs are well separated, and the structure is systematic. The text flows smoothly and is cohesive.	1~5
Presentation(P)	Originality of expression(P1)	The expression is not clichéd and is original. (creative metaphors, witty quotations, literary expressions, etc.)	1~5
	Appropriateness of expression(P2)	The expression is easy to understand and concise. The author uses accurate and objective words that fit the grammar.	1~5
Author's voice(V)	Perspective and personality(V)	The author's own new perspective on the topic is revealed. The author's personality is revealed. (in writing style, etc.)	1~5
Reader's response(R)	Fun and persuasiveness(R)	The writing is fun and interesting. The author's argument is persuasive and the reader is impressed.	1~5
Creativity score			Average

그림 2는 원천데이터 예시이고, 해당 에세이를 표 2의 창의성 평가 지표에 따라 평가한 것이 표 3이다. 각 평가 요소의 최종 점수는 평가자 1, 평가자 2, 평가자 3이 매긴 점수의 평균이며, 최종적으로 해당 에세이의 창의성 점수는 9가지 평가 요소의 평균 점수이다.

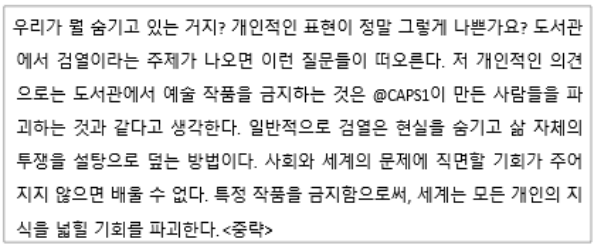

우리가 뭘 숨기고 있는 거지? 개인적인 표현이 정말 그렇게 나쁜가요? 도서관에서 검열이라는 주제가 나오면 이런 질문들이 떠오른다. 저 개인적인 의견으로는 도서관에서 예술 작품을 금지하는 것은 @CAPS1이 만든 사람들을 파괴하는 것과 같다고 생각한다. 일반적으로 검열은 현실을 숨기고 삶 자체의 투쟁을 설탕으로 덮는 방법이다. 사회와 세계의 문제에 직면할 기회가 주어지지 않으면 배울 수 없다. 특정 작품을 금지함으로써, 세계는 모든 개인의 지식을 넓힐 기회를 파괴한다.<중략>

그림 2. 원천데이터 예시
Fig. 2. Example of source data

표 3. 창의성 평가 예시
Table 3. Example of creativity assessment

Category	Evaluation index	Evaluator 1	Evaluator 2	Evaluator 3	Average
C	C1	5	4	5	5
	C2	5	5	5	5
	C3	5	4	5	5
O	O1	5	3	3	4
	O2	5	4	4	4
P	P1	5	3	4	4
	P2	5	4	4	4
V	V	5	4	4	4
R	R	5	4	5	5
Creativity score		5	4	4	4

표 4는 Kendall 상관계수(KROCC, Kendall Rank-Order Correlation Coefficient)[17]를 사용하여 평가자 간 평가 점수 일치도를 측정한 결과이다. Kendall 상관계수는 1에 가까울수록 세 명의 평가자가 같은 평가를 했다는 것을 의미한다. 평가자 간 점수 일치도는 표 1에서의 전체 에세이, 주제별 에세이, 종류별 에세이로 측정하였다. Kendall 점수는 모두 0.8 이상으로 매우 강한 상관관계를 보였다.

표 5와 표 6은 각각 범주별 및 평가 요소별에 대한 평가자 간 Kendall 점수 결과이다.

표 4. 평가자 간 평균 Kendall 점수
Table 4. Average Kendall score between evaluators

Prompt	Average Kendall score
All	0.827
1	0.823
2	0.807
3	0.887
4	0.824
5	0.906
6	0.882
7	0.828
Editorial article	0.824
Information article	0.859

표 5. 범주별 평가자 간 평균 Kendall 점수
Table 5. Average Kendall score between evaluators per category

Prompt	C	O	P	V	R
All	0.844	0.824	0.817	0.807	0.817
1	0.840	0.819	0.812	0.808	0.811
2	0.819	0.789	0.795	0.757	0.859
3	0.892	0.885	0.860	0.867	0.939
4	0.825	0.840	0.856	0.902	0.459
5	0.947	0.912	0.904	0.923	0.777
6	0.914	0.871	0.904	0.775	0.826
7	0.860	0.843	0.796	0.738	0.823
Editorial article	0.841	0.821	0.813	0.806	0.812
Information article	0.879	0.859	0.857	0.801	0.849

표 6. 평가 요소별 평가자 간 평균 Kendall 점수
Table 6. Average Kendall score between evaluators per evaluation index

Prompt	C1	C2	C3	O1	O2	P1	P2	V	R
All	0.837	0.858	0.829	0.802	0.830	0.819	0.804	0.807	0.817
1	0.839	0.851	0.823	0.797	0.826	0.816	0.798	0.808	0.811
2	0.735	0.842	0.838	0.742	0.807	0.818	0.737	0.757	0.859
3	0.807	0.885	0.917	0.854	0.813	0.801	0.860	0.867	0.939
4	0.505	0.869	0.936	0.805	0.755	0.696	0.755	0.902	0.459
5	0.911	0.911	0.942	0.920	0.901	0.901	0.739	0.923	0.777
6	0.864	0.893	0.928	0.835	0.840	0.881	0.855	0.775	0.826
7	0.844	0.900	0.803	0.861	0.786	0.825	0.769	0.738	0.823
Editorial article	0.839	0.854	0.824	0.801	0.827	0.817	0.799	0.806	0.812
Information article	0.811	0.878	0.887	0.837	0.838	0.850	0.804	0.801	0.849

표 7. 데이터베이스 스키마의 예시
Table 7. Example of MongoDB database schema

<table>
<tr><th>Key</th><th>Value</th></tr>
<tr><td>ID</td><td>1</td></tr>
<tr><td>English</td><td><table><tr><th>Key</th><th>Value</th></tr><tr><td>s1</td><td>Sentence</td></tr><tr><td>s2</td><td>Sentence</td></tr><tr><td>s3</td><td>Sentence</td></tr><tr><td>...</td><td>...</td></tr></table></td></tr>
<tr><td>Korean</td><td><table><tr><th>Key</th><th>Value</th></tr><tr><td>s1</td><td>Sentence</td></tr><tr><td>s2</td><td>Sentence</td></tr><tr><td>s3</td><td>Sentence</td></tr><tr><td>...</td><td>...</td></tr></table></td></tr>
<tr><td>Essay score</td><td>3</td></tr>
<tr><td>Worker</td><td>Daegon Yu</td></tr>
<tr><td>Date</td><td>2022.11.13</td></tr>
<tr><td>Source</td><td>ASAP</td></tr>
<tr><td>Type</td><td>Information or editorial article</td></tr>
<tr><td>Topic</td><td>My success story</td></tr>
<tr><td>Creativity score</td><td><table><tr><th>Key</th><th>Value</th></tr><tr><td>Evaluator1</td><td><table><tr><th>Key</th><th>Value</th></tr><tr><td>C1</td><td>3</td></tr><tr><td>C2</td><td>2</td></tr><tr><td>C3</td><td>4</td></tr><tr><td>O1</td><td>5</td></tr><tr><td>...</td><td>...</td></tr><tr><td>Ave</td><td>3</td></tr><tr><td>Date</td><td>2023.01.18</td></tr></table></td></tr><tr><td>Evaluator2</td><td><table><tr><th>Key</th><th>Value</th></tr><tr><td>C1</td><td>3</td></tr><tr><td>C2</td><td>2</td></tr><tr><td>C3</td><td>4</td></tr><tr><td>O1</td><td>5</td></tr><tr><td>...</td><td>...</td></tr><tr><td>Ave</td><td>3</td></tr><tr><td>Date</td><td>2023.01.18</td></tr></table></td></tr><tr><td>Evaluator3</td><td><table><tr><th>Key</th><th>Value</th></tr><tr><td>C1</td><td>3</td></tr><tr><td>C2</td><td>2</td></tr><tr><td>C3</td><td>4</td></tr><tr><td>O1</td><td>5</td></tr><tr><td>...</td><td>...</td></tr><tr><td>Ave</td><td>3</td></tr><tr><td>Date</td><td>2023.01.18</td></tr></table></td></tr><tr><td>Total score</td><td>3</td></tr></table></td></tr>
</table>

3.3 데이터베이스 구축 및 학습데이터 레이블링 시스템

창의적인 에세이 학습데이터를 생성하고 MongoDB에 저장하였다. 표 7은 데이터베이스 스키마의 예시이다. 에세이는 원시데이터(영문)와 원천데이터(한글)를 함께 저장하였다. 에세이 점수는 기존 ASAP에 레이블링 된 점수이며, 작업자는 영문 에세이 번역 및 검수 작업을 한 사람을 의미한다. 그림 3은 데이터베이스와 연동된 레이블링 시스템 화면이다.

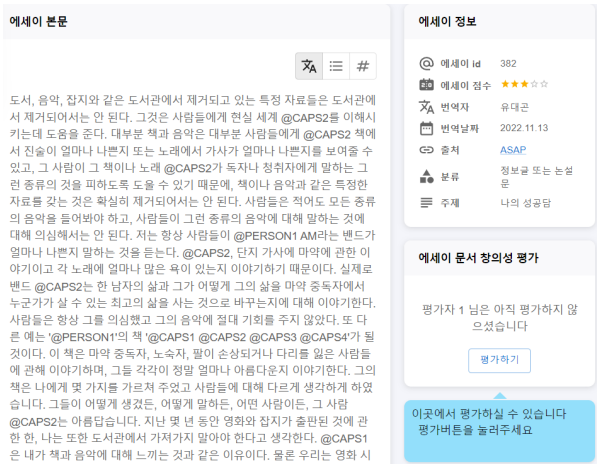


그림 3. 레이블링 시스템 화면
Fig. 3. Screenshot image of labeling system

그림 4~8은 시스템의 대표적인 5가지 기능을 보여준다. 평가자는 번역 기능을 통해 한국어 번역 및 수정 과정에서 간과한 부분을 영어 원문을 보면서 자연스러운 문장으로 수정할 수 있고, 문장 분할 기능과 익명화 처리 기능을 통해 에세이를 문장 단위로 구분하여 가독성을 높일 수 있으며, 개인정보 보호로 인한 익명화의 표기를 헷갈리지 않고 수월하게 평가를 진행할 수 있다. 평가자는 에세이를 읽은 후, 9가지 평가 요소에 대한 평가를 간편하게 마우스 클릭으로 진행할 수 있다. 평가가 완료되면 결과는 자동으로 데이터베이스에 저장되고, 에세이 변환 기능을 통해 다음 에세이 평가를 신속하게 진행할 수 있다.

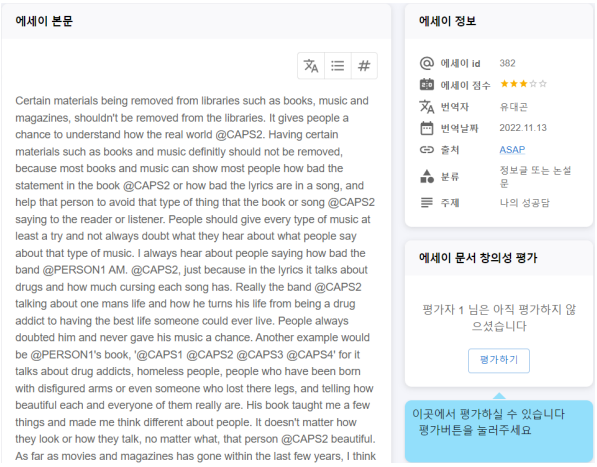


그림 4. 번역 기능 화면
Fig. 4. Screenshot image of translation

그림 5. 문장 분할 기능 화면
Fig. 5. Screenshot image of sentence segmentation

그림 6. 익명화 처리 기능 화면
Fig. 6. Screenshot image of anonymization

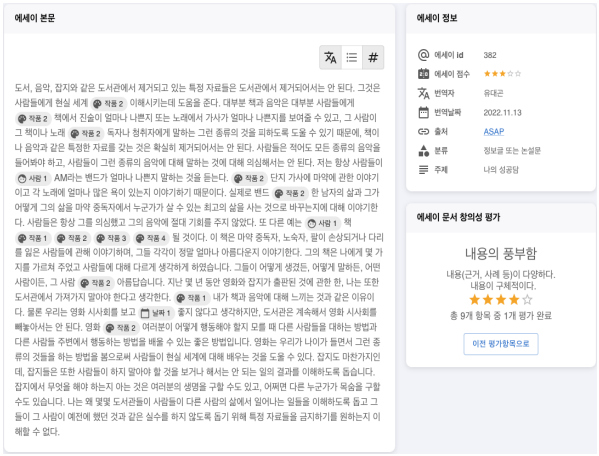


그림 7. 평가 기능 화면
Fig. 7. Screenshot image of evaluation

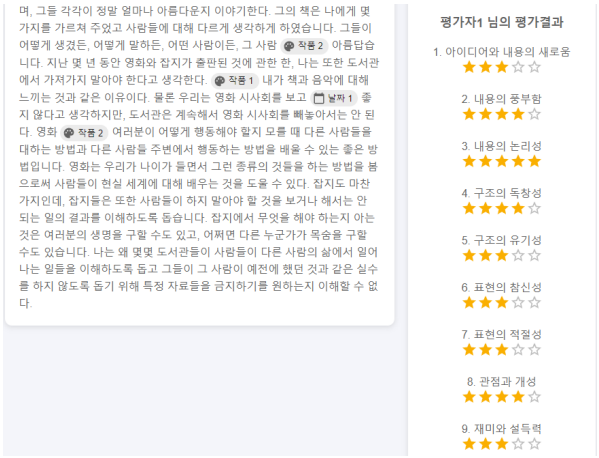


그림 8. 에세이 평가 결과 화면
Fig. 8. Screenshot image of essay evaluation result

전통적인 에세이 평가방식은 두 개의 엑셀 시트(에세이가 저장된 시트와 평가 점수를 기록할 시트)를 번갈아 보면서 평가를 했기 때문에, 상당히 불편하였다. 본 논문에서 제안된 레이블링 시스템은 앞에서 언급한 기능을 통해, 평가자는 더 간편하고 효과적으로 에세이를 평가할 수 있으며, 시간을 절약할 수 있다. 제안한 레이블링 시스템의 정량적 평가를 위해 [18]의 시스템 사용성 척도(SUS, System Usability Scale) 기반 설문조사를 수행하였다. 설문지 내용은 부록 1을 통해 확인할 수 있다. 30명의 응답자가 각 문항에 대해 1점에서 5점으로 답변하였다. 그림 9는 기존 시스템과 제안 시스템에 대한 시스템 사용성 척도 기반의 설문지 문항별 평균값이다. 제안 시스템이 기존 시스템보다 우수하다는 것을 검증하기 위해 모든 문항의 평균값을 활용하여 기존의 평가 시스템과 레이블링 시스템 간의 t-검정(t-test)을 수행하였다. 귀무가설은 두 시스템 간 평균값의 차이가 없다고 가정하고, 대립가설은 두 시스템 간 평균값의 차이가 있다고 가정한다.

귀무가설: $H_0 : \mu_1 = \mu_2$
대립가설: $H_1 : \mu_1 \neq \mu_2$

부록 1. 시스템 사용성 척도 설문지
Appendix 1. System usability scale questionnaire

No.	Question	Strongly disagreement	Disagreement	Neutral	Agreement	Strongly agreement
1	Do you want to use this system often to grade your essays?	①	②	③	④	⑤
2	Is this system unnecessarily complicated?	①	②	③	④	⑤
3	Is this system easy to use?	①	②	③	④	⑤
4	Do you need to get support from an administrator to use this system?	①	②	③	④	⑤
5	Do you think that this system integrates its various functions well?	①	②	③	④	⑤
6	Are there many inconsistencies in this system?	①	②	③	④	⑤
7	Do you think users will be able to quickly learn how to use this system?	①	②	③	④	⑤
8	Is this system very inconvenient to use?	①	②	③	④	⑤
9	Do you feel confident in evaluating your essays using this system?	①	②	③	④	⑤
10	Do you need any background knowledge before using this system?	①	②	③	④	⑤

이렇게 설정된 가설을 기반으로 제안 시스템의 성능이 통계적으로 유의미한지를 검증하기 위해 양측검정을 수행하였다. 이때, 주어진 데이터를 사용하여 귀무가설을 기각할지 아니면 채택할지를 결정하는데 귀무가설이 기각되기 위한 유의수준에 따른 기각역(Critical region)을 결정해야 한다. 이에 대해 95%의 신뢰도를 기준으로 하여 유의수준을 0.05 값으로 결정하였다. 따라서 t-검정을 통해 나온 t-값(t-value)이 기각역에 존재하여 유의확률(p-value)이 0.05보다 작으면 귀무가설은 기각된다.

표 8은 t-검정 결과이다. S_1은 제안한 시스템이고, S_2는 기존 시스템이다. t-Score는 통계학에서 중요한 개념 중 하나로, 주로 가설검정 및 표본 평균의 신뢰구간을 계산하는 데 사용된다. t-Score는 표본 데이터의 평균을 모집단의 평균과 비교하여 모집단에서 표본을 추출한 경우 관찰된 차이가 우연에 의한 것인지 통계적으로 판단하는 데 도움을 준다. t-Score의 절댓값이 클수록 표본 평균과 모집단 평균 간의 차이가 크다는 것을 나타낸다.

결과적으로 t-Score는 9.9350으로, 유의수준이 0.05이고 자유도가 18일 때의 값인 2.101보다 크므로 기각역에 존재한다. 또한, 유의확률(p-value)이 9.873e-09로 유의수준 0.05보다 낮으므로 귀무가설은 기각되고 대립가설을 받아들인다. 즉, 제안된 레이블링 시스템은 기존 평가 시스템보다 우수하다고 말할 수 있다.

Ⅳ. 학습데이터 통계

이번 절에서는 구축한 학습데이터의 통계에 관해 다룬다. 표 9~11은 각각 에세이 창의성 점수의 평균과 표준편차, 범주별, 평가 요소별 창의성 점수의 평균과 표준편차이다.

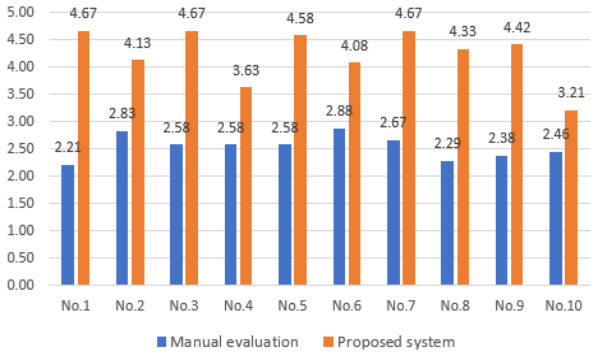


그림 9. 기존 및 제안 시스템의 문항별 평균값
Fig. 9. Average scores per question in SUS

표 9. 창의성 점수의 평균과 표준편차
Table 9. Mean and standard deviation of creativity scores

Prompt	Mean	Standard
All	2.85	0.72
1	2.81	0.72
2	3.25	0.64
3	3.83	0.42
4	3.69	0.34
5	3.43	0.49
6	3.32	0.49
7	3.25	0.53
Editorial article	3.03	0.62
Information article	3.37	0.56

표 10. 범주별 창의성 점수의 평균과 표준편차
Table 10. Mean and standard deviation of creativity scores per category

Prompt	C		O		P		V		R	
	avg	std	avg	std	avg	std	avg	std	avg	std
All	2.85	0.90	2.81	0.88	2.82	0.84	2.87	0.84	2.99	0.85
1	2.81	0.90	2.77	0.88	2.78	0.83	2.83	0.84	2.94	0.84
2	3.05	0.83	3.08	0.80	3.18	0.79	3.74	0.83	3.32	0.94
3	3.81	0.70	3.78	0.55	3.75	0.62	3.88	0.81	4.13	0.62
4	3.56	0.62	4.00	0.60	3.58	0.79	3.17	0.75	4.17	0.41
5	3.48	0.70	3.39	0.78	3.33	0.91	3.11	0.78	3.89	0.60
6	3.41	0.89	3.12	0.97	3.20	0.93	3.30	0.73	3.67	0.74
7	3.26	0.75	3.33	0.66	3.27	0.64	2.89	0.57	3.33	0.67
Editorial article	2.83	0.90	2.79	0.88	2.79	0.83	2.86	0.71	3.13	0.73
Information article	3.35	0.85	3.28	0.87	3.33	0.85	3.46	0.81	3.63	0.82

표 8. t-검정 결과
Table 8. Result of t-test

S_1			S_2			t-Score	p-value
N	Mean	Std	N	Mean	Std		
10	4.2375	0.493	10	2.546	0.2174	9.9350 > α_18, 0.05 = 2.101	9.873e-09

표 11. 평가 요소별 창의성 점수의 평균과 표준편차
Table 11. Mean and standard deviation of creativity scores per evaluation index

Prompt	C1		C2		C3		O1		O2		P1		P2		V		R	
	avg	std	avg	std	avg	std	avg	std	avg	std	avg	std	avg	std	avg	std	avg	std
All	2.68	0.86	2.94	0.94	2.95	0.87	2.61	0.86	3.01	0.86	2.67	0.84	2.97	0.81	2.87	0.84	2.99	0.85
1	2.66	0.88	2.88	0.92	2.90	0.87	2.58	0.86	2.95	0.86	2.64	0.85	2.91	0.79	2.83	0.84	2.94	0.84
2	2.68	0.59	3.54	1.17	3.20	0.88	2.73	0.63	3.58	0.62	2.80	0.62	3.66	0.62	3.74	0.83	3.14	0.91
3	3.31	0.70	4.25	0.58	3.88	0.50	3.50	0.63	4.06	0.25	3.50	0.63	4.00	0.52	3.88	0.81	4.06	0.44
4	3.17	0.41	4.00	0.63	3.50	0.55	3.67	0.52	4.33	0.52	3.00	0.63	4.17	0.41	3.17	0.75	4.33	0.52
5	3.00	0.71	3.89	0.60	3.56	0.53	3.00	0.71	3.78	0.67	2.67	0.71	4.00	0.50	3.11	0.78	3.67	0.71
6	2.73	0.63	3.88	0.82	3.64	0.78	2.58	0.79	3.67	0.82	2.61	0.79	3.79	0.65	3.30	0.73	3.67	0.74
7	2.89	0.63	3.49	0.80	3.41	0.69	3.11	0.63	3.56	0.62	3.17	0.64	3.37	0.63	2.89	0.57	3.27	0.83
Editorial article	2.77	0.75	3.18	0.86	3.16	0.78	2.85	0.74	3.25	0.74	2.91	0.74	3.14	0.71	2.86	0.71	3.13	0.73
Information article	2.90	0.65	3.68	0.92	3.48	0.76	2.87	0.80	3.69	0.73	2.86	0.77	3.79	0.65	3.46	0.81	3.63	0.82

V. 실 험

이번 절에서는 구축한 3,766개의 학습데이터를 검증하기 위해 진행한 실험에 관해 다룬다. 1,800개의 영어 에세이는 BERT-base 모델을 사용하였고, 1,966개의 한글 에세이는 BERT-base-multilingual을 사용하였다. 사전 훈련된 BERT 모델 학습 시 미세조정 단계에서 BertForSequenceClassification을 사용하였다. 이는 텍스트 시퀀스를 입력으로 받아, 해당 시퀀스를 특정 클래스 또는 레이블로 분류하는 작업에 사용된다. 또한, 추가로 사전 훈련된 GPT-2 모델과 GPT2ForSequenceClassification 모델로 BERT와 비교 실험하였다. 학습데이터의 훈련, 검증, 테스트 비율을 8:1:1로 나누었다.

표 12는 모델별로 에세이 창의성 점수(1~5점)의 분류에 대해 정확도, 정밀도, 재현율, F1 점수를 측정한 결과이다. 실험 결과, 정확도는 영어 에세이에 대해 분류한 BERT 모델이 67.8%로 가장 높았고, 정밀도, 재현율, F1 점수는 영어 에세이에 대해 분류한 GPT-2 모델이 각각 72.3%, 49.7%, 58.9%로 가장 높았다. 1~5점으로 클래스 개수가 5개인 다중 분류 문제인 것을 고려하면 비교적 준수한 실험 결과로 보이나 향후 성능이 개선된 모델의 연구가 필요하다. 클래스 비율을 살펴본 결과, 3점으로 평가된 에세이가 전체의 54%를 차지하였고, 1점과 5점으로 평가된 에세이는 각각 0.047%, 0.021%로 매우 낮은 비율이었다. 부족한 1점, 5점의 에세이 학습데이터를 더 구축하여 클래스 불균형 문제를 보완한다면 모델의 성능을 높일 수 있을 것이다. 또한, BERT와 GPT-2는 각각 최대 512 토큰과 1024 토큰까지만 처리할 수 있는 한계가 있다.

실험에 사용된 에세이 대부분은 1,024 토큰을 초과한다. 따라서 초과한 토큰 수 만큼 에세이의 뒷부분이 잘려서 학습되어 모델이 모든 문맥을 고려하지 못하는 한계점이 있다.

표 12. 창의성 점수 예측 모델에 대한 정확도, 정밀도, 재현율, F1 점수
Table 12. Accuracy, precision, recall, and F1 score of essay creativity score prediction models

Model		Accuracy	Precision	Recall	F1 score
BERT	ENG	0.678	0.554	0.418	0.476
	KOR	0.629	0.465	0.445	0.455
GPT-2	ENG	0.667	0.723	0.497	0.589
	KOR	0.584	0.336	0.333	0.334

VI. 결론 및 향후 연구

본 논문에서는 인공지능 모델 학습을 위한 에세이 창의성 학습데이터를 구축하는 방안을 처음으로 제안하였으며, 9가지 창의성 평가 요소를 고려한 학습데이터를 구축할 수 있는 웹 기반 레이블링 시스템을 제안하였다. 데모 시스템은 dilab.kunsan.ac.kr:3000/work/1/tool/1에서 직접 확인할 수 있다.

향후 연구로는 본 논문에서 제안한 레이블링 시스템을 통해 AIHUB에 있는 에세이 데이터를 가지고 에세이 창의성 평가 학습데이터를 새로 구축할 것이다. 그 이유는 AIHUB 데이터 세트는 주로 문법, 철자 오류, 논리적 일관성 등을 평가하는 데 사용되어 창의성을 평가하는 데에 한계가 있다. 또한, 표 12의 결과처럼, 현재 최고 성능을 보이는 BERT와 GPT-2 모델을 사용해도 성능은 그리 높지 않다. 학습데이터를 추가로 구축하여 앞서 언급한 클래스 불균형 문제를 보완하고 4,096 토큰 이상의 긴 문서 처리가 가능한 Longformer 모델을 사용하여 정확도를 향상할 수 있는 인공지능 모델을 개발할 것이다.

Acknowledgement

본 논문의 연구를 수행하는 데 많은 도움을 주신 퇴계원고등학교 유지연 국어 선생님과 한림연예예술고등학교 윤종대 국어 선생님께 감사를 드립니다.

References

[1] M. W. Noh and R. Y. Kim, "A study on Analysis of Leaner Responses in Web board-based Reading Discussions", Korea Reading Association, No. 20, pp. 171-199, 2008.

[2] E. P. Torrance, "The Torrance Tests of Creative Thinking: Norms-Technical Manual Princeton", NJ: Personal Press, 1974.

[3] R. J. Sternberg and T. I. Lubart, "The Concept of Creativity: Prospects and Paradigms", Handbook of Creativity, pp. 3-15, Cambridge University Press, 1999.

[4] J. A. Plucker, R. A. Beghetto, and G. T. Dow, "Why Isn't Creativity More Important to Educational Psychologists? Potentials, Pitfalls, and Future Directions in Creativity Research", Educational Psychologist, Vol. 39, No. 2, pp. 83-96, Jun. 2004. https://doi.org/10.1207/s15326985ep3902_1.

[5] S. Y. Oh, "Study on the Aspects of Creativity Scoring in the Writing Assessment of Information Text", The Korean Language And Culture Education Society, No. 28, pp. 59-93, May 2013.

[6] T. Ghosal, V. Edithal, A. Ekbal, P. Bhattacharyya, G. Tsatsaronis, and S. Chivukula, "Novelty Goes Deep. A Deep Neural Solution To Document Level Novelty Detection", 27th International Conference on Computational Linguistics (COLING), Santa Fe, New Mexico, USA, pp. 2802-2813, Aug. 2018.

[7] R. K. Amplayo, S. Hwang, and M. Song, "Evaluating Research Novelty Detection: Counterfactual Approaches", Thirteenth Workshop on Graph-based Methods for Natural Language Processing, Hong Kong, pp. 124-133, Nov. 2019. http://dx.doi.org/10.18653/v1/D19-5315.

[8] B. Bhattarai, O. Granmo, and L. Jiao, "Measuring the Novelty of Natural Language Text using the Conjunctive Clauses of a Tsetlin Machine Text Classifier", arXiv preprint arXiv:2011.08755, Nov. 2020. https://doi.org/10.48550/arXiv.2011.08755.

[9] S. Doboli, J. Kenworthy, P. Paulus, A. Minai, and A. Doboli, "A Cognitive Inspired Method for Assessing Novelty of Short-text Ideas", International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, pp. 1-8, Jul. 2020. https://doi.org/10.1109/IJCNN48605.2020.9206788.

[10] G. Liang, B. On, D. Jeong, A. A. Heidari, H. Kim, G. Choi, Y. Shi, Q. Chen, and H. Chen, "A Text GAN Framework for Creative Essay Recommendation", Knowledge-Based Systems, Vol. 232, Nov. 2021. https://doi.org/10.1016/j.knosys.2021.107501.

[11] Kaggle, "The Hewlett Foundation: Automated Essay Scoring", https://www.kaggle.com/competitions/asap-aes/data [accessed: Jun. 21, 2022]

[12] Y. Wu, et al., "Google's neural machine translation system: Bridging the gap between human and machine translation", arXiv preprint arXiv:1609.08144, Oct. 2016. https://doi.org/10.48550/arXiv.1609.08144.

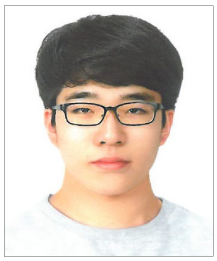
[13] J. Y. Park, "How do thoughts become writing", Sam & Parkers, Mar. 2020.
[14] J. R. Han, "Development of Standards for Creativity Assessment of Narrative Text and Persuasive Text", Graduate School of Korea National University of Education, Korean Language Education Major, master's thesis, 2015.
[15] S. Y. Lee, "Recent Trends and Challenges in Reading Research", Center for Korean Literature & Language Education in Korea University, No. 10, pp. 311-340, 2011.
[16] J. H. Park, "A Study on the concept of 'Creativity' in Korean Language Education with Regard to Writing Evaluation", The Society Of Korean Language And Literature Education, No. 20, pp. 383-406, Aug. 2004.
[17] M. G. Kendall, "A New Measure of Rank Correlation", Biometrika, Vol. 30, No. 1/2, pp. 81-93, Jun. 1938. https://doi.org/10.2307/2332226.
[18] J. Brooke, "SUS: A 'Quick and Dirty' Usability Scale", Usability Evaluation in Industry, Jun. 1996. https://doi.org/10.1201/9781498710411-35.

저자소개

김 덕 기 (Deokgi Kim)

2016년 3월 ~ 현재 : 군산대학교 소프트웨어학부 학사과정
관심 분야 : 자연어처리, 인공지능

한 상 우 (Sangwoo Han)



2018년 3월 ~ 현재 : 군산대학교 소프트웨어학부 학사과정
관심 분야 : 자연어처리, 인공지능

박 승 혁 (Seunghyeok Park)

2018년 3월 ~ 현재 : 군산대학교 소프트웨어학부 학사과정
관심 분야 : 자연어처리, 인공지능

김 용 연 (Yougyeon Kim)

2017년 3월 ~ 현재 : 군산대학교 소프트웨어학부 학사과정
관심 분야 : 자연어처리, 인공지능

박 상 현 (Sanghyun Park)

2018년 3월 ~ 현재 : 군산대학교 소프트웨어학부 학사과정
관심 분야 : 자연어처리, 인공지능, 강화학습

조 준 영 (Junyoung Jo)

2019년 3월 ~ 현재 : 군산대학교 소프트웨어학부 학사과정
관심 분야 : 자연어처리, 인공지능, 강화학습

최 서 인 (Seoin Choi)

2021년 3월 ~ 현재 : 군산대학교 소프트웨어학부 학사과정
관심 분야 : 자연어처리, 인공지능

이 유 빈 (Youbin Lee)

2023년 2월 : 군산대학교
소프트웨어학과(학사)
관심 분야 : 자연어처리, 인공지능

유 대 곤 (Daegon Yu)

2023년 2월 : 군산대학교
소프트웨어학과(학사)
2023년 4월 ~ 현재 : ㈜애니파이브
연구원
관심 분야 : 자연어처리, 인공지능

온 병 원 (Byung-Won On)

2007년 : 펜실베이니아주립대학교
컴퓨터공학과 박사
2008년 ~ 2009년 :
브리티시컬럼비아대학교
컴퓨터과학과 박사후연구원
2010년 : 차세대디지털과학센터
선임연구원
2011년 ~ 2014년 : 차세대융합기술연구원 선임연구원
2014년 ~ 현재 : 군산대학교 소프트웨어학부 교수
관심 분야 : 자연어처리, 인공지능, 강화학습